FREQUENT PATTERN MINING ON ONLINE JUDGE EDUCATION WEB LOG DATA USING ECLAT ALGORITHM

POE MYAT ZIN

M.C.Sc. **JUNE 2022**

FREQUENT PATTERN MINING ON ONLINE JUDGE EDUCATION WEB LOG DATA USING ECLAT ALGORITHM

By

Poe Myat Zin

B.C.Sc

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Master of Computer Science

(M.C.Sc.)

University of Computer Studies, Yangon
June 2022

ACKNOWLEDGEMENTS

I wish to express my sincere appreciation to **Prof. Dr.Mie Mie Khin**, the Rector of the University of Computer Studies, Yangon, for her kind permission to develop this thesis.

My sincere thanks and regards go to **Dr. Si Si Mar Win**, Professor, Faculty of Computer Science, University of Computer Studies, Yangon, for her kind management throughout the completion of this thesis.

My sincere thanks to **Dr. Tin Zar Thaw**, Professor, Faculty of Computer Science, University of Computer Studies, Yangon, for her kind management throughout the completion of this thesis.

I would like to express my heartfelt thanks to my supervisor, **Daw Aye Aye Maw,** Associate Professor, Faculty of Computer Science, University of Computer Studies, Yangon, for her encouragement, patient, and invaluable supervision in compiling the materials for my thesis.

I also deeply thank **Daw Hnin Yee Aung**, Lecturer, Department of English, University of Computer Studies, Yangon for editing my thesis from the language point of view.

Finally, I especially thank my parents, all of my friends and their suggestions, support and generous help rendered to me during the development of thesis. Their love and concerned encouragement have strengthened me through the studies.

Moreover, I would like to thank all the staffs, teachers from the University of Computer Studies, Yangon for their support.

ABSTRACT

The World Wide Web is a popular and interactive tool for spreading information today that is expanding rapidly. Web mining is the process of extracting patterns and implicit data from artifacts or activities associated to the World Wide Web in order to find relevant and potentially helpful patterns. This extracted data can also be utilized to enhance web personalization, fraud detection, future prediction accessed by the user, user profiling, and understanding of user web activity. Upcoming page prediction is one example of how this information can be used to improve web usage mining. The proposed system consists of three phases which are data pre-processing, pattern discovery, and pattern analysis. Raw web log data may contain noise and impurities. By using some data preprocessing techniques that noise will be removed. Data preprocessing phase is the most important one because it makes the data with good quality. In Pattern discovery phase, the users' navigational pattern and rules are extracted by using association rule algorithm. In this thesis, first preprocessing can be done with data cleaning, user identification, page identification and session identification. The objective of this thesis is to identify the frequent pattern from Online Judge web log data of web server of the University by using the ECLAT algorithm. We modified a very efficient ECLAT algorithm for matching interesting new patterns and even used support and confidence to calculate interesting pattern measures.

TABLE OF CONTENTS

Pages

LIST OF FIGURES

Pages

LIST OF TABLES

Pages

CHAPTER 1
INTRODUCTION

The World Wide Web is the leading data source for millions of people around the world to access information from vast amounts of available data such as advertising, consumer information, e-commerce, education and finance administration, government, news and many other information services. While searching for specific information on the web, It is important that the user retrieves the information in a short time. Web suggestion models make it easier for users to access a website while browsing. Prediction patterns play an important role in e-commerce and advertising on specific pages of a commercial website. Web traffic forecasting is useful for customizing your web content to send personalized content to specific users. Web data mining is a powerful technique in the field of data mining and emerging research, where different techniques have been used to solve various problems related to analyzing user web usage patterns available in web servers. Implement web log mining to identify real-world issues such as frequently visited, most visited pages, interesting users, and emerging patterns of user behavior.

Web mining is the process of extracting web content; it is a data mining technology that automatically detects and retrieves information by analyzing web structure and web usage. Applying web content mining, meaningful data can be obtained from web document content. Finding structure-related information from the web is a process called web structure mining. The component and link structure of a web site is analyzed using web structure mining. This is one of web usage mining technique prediction the normal user’s browsing behaviors. Web mining is used to find interesting patterns in web log data. It is possible to identify associations between groups of users who have a similar interest by using association rules to find the pages that are frequently visited together, even if they are not directly related. By providing linkages between pages that are frequently visited together, this information is essential, for instance, to restructure websites. Association rule is a model that identifies specific types of data associations. The retail sales community can identify products that are frequently bought together by using these associations. In Web usage mining several data mining techniques can be used. In a web environment, association rules are often applied to former user session data seen in HTTP server log files. Web sessions are collected without the input of any users, and they also organized user behavior when exploring a web site. Web sessions can therefore be viewed as a significant source of data on users. Web Usage Mining is a subset of data mining that has got a lot of attention in the last year. Commercial companies and technical researchers have developed a wide range of tools that perform various data mining algorithms on web server document files to track user behavior on a specific Web site. This type of inquiry on the Website can provide usable information to tailor the user's needs.

1.1 Objectives of the Thesis

The primary objective of this thesis is to identify frequent patterns by using association rule method on web log data. The rest of other objectives are:

- To study the concepts of Web Usage Mining how is used to extract the web user’s behavior
- To learn the most important phase the steps by steps of preprocessing of web log data
- To generate the useful association rules from the frequent itemset
- To provide the effective frequent pattern mining algorithm on web log data
- To analyze the nature of Online Judge Web Log Data and to study how to remove irrelevant data or noise

1.2 Motivation of the Thesis

The Internet is one of the greatest inventions and has become the most popular resource for disseminating and enhancing information as well as extracting useful information. As the Internet became more and more popular, Web servers automatically collect large amounts of data and store it in log files. Analyzing server access data can provide clear and useful information. Website is a group of web pages. Text, pictures, and videos can all be found on web pages. Through hyperlinks, which allow for navigation, they are connected. The amount of log files is growing daily as a result of the increased usage of websites.

Web mining is the use of various data mining techniques to extract data from web-based services and documents. In this thesis, we address the problem of Web usage mining, which is the mining of user frequent patterns from one or more web servers in order to find relationships between data stored, with a focus on interesting new patterns. By enhancing website design, analysis of user behavior patterns and interests increases website performance. An efficient ECLAT Algorithm is adapted for matching interesting new patterns and applied support and confidence to calculate the measures of interesting patterns.

1.3 Related Works

In this paper, they addressed the problem of Web usage mining, i.e. finding relationships between data stored, user frequent patterns from one or more Web servers were mined and paired particular attention to the interesting new patterns. Apriori algorithm was adapted for matching interesting new patterns and measuring of interesting patterns by applying support and confidence, to this particular context. [8]. Web mining describes the process of Web data mining in detail: source data collection, data pre-processing, pattern discovery, pattern analysis and cluster analysis. Servers are able to collect and store huge amount of data by the help of advanced information technologies. To study the customers' behavior was the main purpose of this paper and they used the Web mining techniques and tested its application in e-commerce [6]. To discover the hidden knowledge and to identify the behavior of the user on the web, this paper used the web data sources. The web access log file collected from the organization was used to examine the user pattern by using the updated Web Log Expert application. This enhanced tool attempted to carry out web mining in a domain-agnostic manner. There were three parts to this algorithm: 1. Give an input item, extract a set of IP addresses and visor lists and rank them by comparison. 2. Identify and summarize the competitive data that specifies the organization's strength and 3. Identify and summarize the domains in which the specified entity participates. The Web Log Expert tool had been used to implement the entire analytic process. The findings of the experiment made it easier to traverse the website and improve its design architecture [9]. In this paper, they would learn how to use R algorithms to find common patterns, association rules, and correlation rules. Then, using benchmark data, they would assess all of these ways to see how fascinating the common patterns and rules are. They came to the conclusion that the Apriori algorithm, which was the first efficient mining method for mining common patterns, is the source of many variants [2]. This system aimed to utilize an intelligent technique to deliver personalized web service for accessing related web pages more rapidly and effectively, so that it could be determined which web pages the user is most likely to visit in the future. For forecasting user behavior, this system incorporated two intelligence algorithms: FP Growth and Eclat. These methods solved the existing system's time and space problems. In addition to the frequent pages pattern, Direct and Indirect Association Rules were developed, and a ranking is assigned to pages based on these rules, which would aid the recommendation engine in recommending related search pages [3].

1.4 Organization of the Thesis

This thesis is mainly composed of five chapters. The first chapter introduces the thesis and motivation, as well as the scope and objectives. Chapter 2 describes the literature of association rule methods, highlighting some systems that provide important knowledge and background history for understanding the subsequent chapters. Chapter 3 discusses the system's theoretical foundation as well as the Preprocessing phases. Chapter 4 provides an overview of the system's design and implementation using node.js. Chapter 5 discusses the conclusion and future research directions, as well as problems for future research.

CHAPTER 2

BACKGROUND THEORY

This chapter presents about of web mining with its type and also describe web usage mining technique, association rule mining that is used for web page prediction. First of all, this chapter identifies the meaning of web mining, which is useful for discovering and retrieving interesting patterns from large sets of data. Secondly, it will explain web usage techniques that include preprocessing, data cleaning, user identification, session identification and page identification. Finally, it describes about of association rule mining methods with functionalities and application area.

2.1 Web Mining

Web mining is a subset of data mining that focuses on the World Wide Web as the primary data source, including everything from web content to server logs. The content of the data extracted from the web is text, Structured data and images, such as lists and tables; It is safe to say that information, including video and audio, should be stored on the Web.Web mining has become popular for a number of reasons. The World Wide Web (www) is the best and most significant source for data warehousing and mining. The Web is constantly expanding in size. The source demonstrates that there are more than 10 million publicly accessible Web pages. Additionally, every month the internet has access to almost 6 gigabytes of new data. According to the source, web mining is defined as follows: Web mining is the process of automatically finding and extracting information from documents and Web services using data mining techniques.

Web mining, like traditional data mining, aims to discover and retrieve useful and interesting patterns from large data sets. On web mining, big data acts as data sets. Information, documents, structure, and profile are all components of web data. Web mining is defined by two concepts: process-based and data-driven. According to Web mining data, the web is used to extract knowledge. In general, web mining is used in several steps: data collection, data selection before processing, knowledge discovery, and analysis.

2.1.1 Web content mining

The process of extracting useful information from the content of Web documents is known as web content mining. The contents of a web document correspond to the concepts that the document attempted to convey to users. Text, image, video, sound, or records such as lists and tables can all be included in this content. The technologies used in this policy are largely designed from natural language processing (NLP) and information retrieval.

2.1.2 Web structure mining

This procedure involves using diagram theory to examine a website's nodes and connection structure. This can be used to determine two things: the structure of a website in terms of how it connects to other websites, as well as the document structure of the website itself, or the connections between each page.

2.1.3 Web usage mining

This is the location of the users, the process of extracting patterns and information from server logs to gain insight into users' activity, including how many objects are clicked on the site and how they are active on the site.

Web usage mining is the use of data mining techniques to discover patterns on the Web in order to better understand and meet the user's needs. This type of web mining investigates data relating to web user behavior. It should be noted that there is no clear distinction between web mining groups. For example, Web content mining technologies can access user information in addition to documents.

With the continued growth and proliferation of e-commerce, web services, and web-based information systems, the number of clickstreams and user data that web-based organizations collect in their day-to-day operations has reached astronomical proportions. Such data analysis can assist these organizations in determining the lifetime value of their clients, designing cross-marketing strategies across products and services, evaluating the effectiveness of promotional campaigns, optimizing the functionality of Web-based applications, offering more personalized content to visitors, and implementing effective logical structure for their Web space.This type of analysis primarily involves auto-discovering meaningful patterns and relationships in large collections of semi-structured data. Web-based exploration is the process of automatically discovering and analyzing related data that is collected or released in relation to click flow due to user interactions on one or more websites. The goal is to capture, model, and analyze the behavioral patterns and profiles of website visitors. Forms are usually pages that are frequently visited by a group of users who have a specific need or interest. Represents a collection of objects or resources.

The overall Web usage mining process can be divided into three interdependent stages based on the standard data mining process: data collection and pre-processing, pattern discovery, and pattern analysis. In the pre-production phase, client data is purified and categorized into individual user groups that represent the activities of each user in different visits to the website. In addition to site content or structure, site ontologies can also be used to pre-populate other sources of knowledge, such as se mantic domain knowledge, or to enhance user payment data. In the model search phase, Statistics; Database and machine learning functions are integrated with the normal user behavior and web resources; work to obtain secret forms that reflect summary statistics of sections and users. In the final stages of the process, the models and statistics found are further refined and refined for suggestions engines, Visual aids; It can create aggregate user models that can be used as applications for applications such as web analytics and report generation tools.

2.2 Association Rule Mining

Association Rule Mining; as the name suggests, If / Then statements help to find relationships between seemingly independent related databases or other data repositories that are relatively simple. Most machine learning algorithms are mathematical because they work with numerical data sets. However, the excavation of association rules is not a statistical one. A little more than just a simple calculation is appropriate for categorizing data.

Association rule mining relates to relational databases; Frequent patterns from datasets found in different types of databases, such as transactional databases and other repositories; a procedure intended to monitor relationships or associations.

2.2.1 Measures of the Effectiveness of Association Rules

The strength of an organization's structure is determined by both support and confidence. A rule given in the database is called the frequency of mining or support. The number of times a given rule has been proven true in practice is called the confidence. A rule may be strongly correlated in a dataset because it occurs frequently, but it may appear much less frequently when applied. This is an example of strong support, but not enough confidence.

By contrast, a rule may not be prominent in the dataset, but further investigation shows that it appears frequently. This is an example of high confidence but low support. Using these metrics allows analysts to distinguish causality from correlation and correctly evaluate a given rule.

Association rule: given a set of items I = $\{I_1, I_2, ..., I_n\}$ and database transactions D = $\{t1, t_2, \dots, t_n\}$ where $t_i = \{I_{i1}, I_{i2}, I_{i3}, \dots, I_{ik}\}$ and $I_{ij} \in$ I. An association rule is an implication of the form X => Y where X,Y belongs to I, are set of items called item sets and X ∩ Y = Ø; X is called the antecedent and Y is called consequent of the rule.

- support: the support(S) for an association rule X => Y is the percentage of transactions in the database that consists of XUY.
- confidence: the confidence for an association rule X => Y is the ratio of the number of transactions that contains XUY to the number of transactions that contains X.

Remember that an association rule is an indication of the form XY [sup, conf], where X and Y are item sets, sup is the support of the itemset X Y able to represent the probability that X and Y occur together in a transaction, and conf is the rule's confidence, categorized by sup(XY) / sup(X), representing the conditional probability that Y occurs in a transaction given that X occurs.

The mining of association rules in Web transaction data has many advantages.

For example, a high-confidence rule such as

special-offers/, /products/software/ → shopping-cart/

might provide some indication that a promotional campaign on software products is positively affecting online sales. Such rules can also be used to optimize the structure of the site. For example, if a site does not provide direct linkage between two pages A and B, the discovery of a rule, A → B, would indicates that providing a direct hyperlink from A to B might aid users in finding the intended information. Both association analysis (among products or pageviews) and statistical correlation analysis (generally among customers or visitors) have been used successfully in Web personalization and recommender systems.

2.2.2 Application area of Association Rule Mining

Organizational association rules are used in data science to find relationships and cohesions between data sets. It is best used to interpret data plans from databases, such as relational and commercial databases, which appear to be separate sources of information. The use of association rules is often referred to as "mining associations" or "association rules mining".

Market Basket Analysis

This is the most common example of association mining. Most supermarkets collect data using barcode scanners. This database, also known as the "Marketing Request" database, contains many previous trading records. A record is a list of all items purchased from a customer in a sale.

Medical Diagnosis

The rules of the Medical Diagnostic Association are useful in assisting physicians in treating patients. Diagnosis is not an easy process and contains errors that can lead to unreliable end results. Relational association rule mining can be used to diagnose a variety of causes and symptoms. In addition, this area can be expanded by using learning techniques to determine the relationship between new symptoms and diseases associated with new symptoms.

Association rules can aid in patient diagnosis for doctors. Since many diseases share symptoms, there are numerous factors to take into account while reaching a diagnosis. Doctors can compare symptom associations in the data from previous cases to calculate the conditional probability of a given sickness using association rules and data analysis powered by machine learning. The machine learning model can modify the rules to reflect the most recent information as new diagnoses are made.

Physicians can use machine learning-powered association rules and data analysis to compare symptom associations in previous case data to calculate conditional probabilities for a given disease. Machine learning models can modify the rules to reflect the latest information when conducting new diagnostics.

Retail

Using point-of-sale systems to scan product barcodes, retailers can obtain information about consumers' purchasing habits. To identify the products that are most likely to be bought together, machine learning models can search this data for co-occurrence. The retailer can then use this information to modify its marketing and sales tactics.

User experience (UX) design

Data about user behavior can be gathered by website developers. Then, by researching where people often click and what factors increase their likelihood of interacting with a call to action, for example, they can use the relationships in the data to optimize the website user experience.

Entertainment

Association rules can be used to power content recommendation engines for services like Netflix and Spotify. To suggest content that may be of interest to users, or to gather information in a way that prioritizes what is most interesting to a user; Machine learning models analyze historical user behavior data to find recurring trends.

2.2.3 Association Rule in Data Mining

In data mining, organizational rules are useful for analyzing and predicting user behavior. They are responsible for customer analysis, Marketing basket analysis; Product portfolio; an important part of the catalog design and store layout. Programmers use team rules to build machine-learning programs. Machine Learning is a type of artificial intelligence (AI) that seeks to build more efficient programs without special programming.

A typical example of association rule mining is the relationship between diapers and beer. The example appears to be fictitious, claiming that men who go to the store to buy diapers may also buy beer. The data pointing to this might look like this: A supermarket has 200,000 customer transactions. About 4,000 transactions, or about 2% of total transactions, included the purchase of diapers. About 5,500 transactions (2.75%) included the purchase of beer. Of these, about 3,500 transactions, or 1.75%, included the purchase of diapers and beer. According to the percentage, this large number should be much lower. However, the fact that approximately 87.5% of diaper purchases include beer purchases suggests a link between diapers and beer.

2.2.4 Association Rule Algorithms

Common algorithms that use association rules include AIS, SETM, Apriori, and variations of the latter.

The SETM algorithm also generates frequent itemset while scanning the database, but it considers itemset at the end of the scan. The new candidate itemset is generated in the same way as the AIS algorithm, but the transaction ID of the generated transaction is stored in a sequential data structure along with the candidate itemset. The support count of candidate itemset is calculated at the end of the pass by aggregating the sequential structure.

The Apriori algorithm generates candidate itemset by using only the large itemset from the previous pass. The large itemset of the previous pass are concatenated with themselves to produce all itemset of size one greater. Then delete each generated itemset with a small subset. Candidates are the remaining itemset. Each subset of itemset is often defined as itemset by the Apriori algorithm. The algorithm reduces the number of candidates considered by looking for items larger than the minimum support amount.

Decision trees: Given a database D = {t_1, t_2, …, t_n} where t_i = {t_{i1}, t_{i2}, t_{i3}…, t_{ik}} and the database schema contains the following attributes {A1, An}. Also given set of classes C = {C_1, ..., C_n}. A decision tree is a tree associated with D that has the following properties, each internal node is labelled with an attribute A_i, each arc is labelled with a predicate, and each leaf node is labelled with a class C_i. Decision trees are often used for categorizing and estimating data and for classifying data. Decision trees do have advantages over association rules.

- While association rules on the same target attributes may refer to overlapping subsets, a decision tree divides the dataset.
- A decision tree is guaranteed to have at least 50% prediction accuracy, whereas association rules are disconnected from one another and do not reflect a predictive model of the dataset.

CHAPTER 3

METHODS OF THE PROPOSED SYSTEM

This chapter presents the background theory of web usage mining, preprocessing stage of web log file and association rule mining algorithm. Firstly, this chapter describes about the nature of web server log file data. Secondly, it explains how to clean or remove irrelevant data from these log file. Thirdly, it also describes about step by step of preprocessing which include data cleaning, user identification, page identification and session identification. Finally, it presents about frequent itemset of the Eclat Algorithm.

3.1 Server Log File

Log files are files that contain a record of the actions that have taken place on the web server. Web servers are the computers that serve up web pages. All of the files required to display Web pages on the user computer are stored on the Web server. A website's completeness is formed by the combination of all individual web pages. Images/graphics files and any scripts required for the operation of the website's dynamic elements When a browser requests information from a web server, the server responds with it through HTTP and sends it back to the browser. The files are then converted or formatted by the browser into a page that can be viewed by users. It gets displayed in the browser. Similar to how the server may transfer data to numerous client computers at once, enabling multiple clients to view the same page.

A server log is a log file (or numerous files) that the server automatically creates and maintains, lists all the operations which are carried out. A web server log that keeps track of past page requests is a common illustration. For web server log files, the W3C maintains a standard format (the Common Log Format), however there are also additional proprietary formats. The file ends when more current entries are added. This normally includes the client's IP address, the request date and time, the requested page, the HTTP code, the number of bytes provided, the user agent, and the referrer. This information can either be compiled into a single file or divided into other logs, such as an access log, an error log, and a referrer log. However, the user-specific data is not normally collected by server logs.

Usually, only the webmaster or another administrative person has access to these files on the Internet. Examining traffic patterns by time of day, day of the week, referrer, or user agent may be done using statistical analysis of the server log. Analysis of the web server logs can help with effective website management, proper hosting resources, and fine-tuning of sales efforts.

Analytics on web server logs are done on the values in the log file to determine when, how, and by whom a web server is visited. Reports are typically generated instantly, however, data taken from log files can also be saved in a database, enabling the generation of various reports as needed.

3.1.1 Contents of Log File

Different types of information are stored in log files on different web servers. The log file contains the following basic information:

User name: This identifies who has visited the website. The IP address assigned by the Internet Service Provider is primarily used to identify the user (ISP). This could be a temporary address. As a result, the user's unique identification is lagging. In some websites, user identification is accomplished by obtaining the user profile and authorizing them access to the website via a username and password. The user is uniquely identified in this type of access so that the user's return can also be identified.

Visiting Path: The path taken by the user while visiting the website is referred to as the visiting path. This can be done by directly entering the URL, clicking on a link, or using a search engine.

Path Traversed: It identifies the user's path through the website using the various links.

Time stamp: The amount of time is spent by the user on each web page while browsing the website. It is referred to as the session.

Page last visited: The last page is visited by the user before leaving the website.

Success rate: The number of downloads and copying activities performed by the user can determine the success rate of the website. If any items or software are purchased, the success rate will increase.

User Agent: This is simply the browser through which the user sends the request to the web server. It is simply a string that describes the type and version of browser software that is being used.

URL: The resource is accessed by the user. It may be an HTML page, a CGI program, or a script.

Request type: The method of information transfer is mentioned. Methods such as GET and POST. These are the contents of the log file. This log file information is used in the web usage mining process. According to web usage mining, it mines the most frequently visited websites. The usage would be the most frequently visited website or the website that has been used for a longer period of time. If the log file is examined, the quantitative usage of the website can be determined.

3.2 Using Log File Data in Web Usage Mining

In the proposed system, an Online Judge web log data have been taken of one of the University. There are so many log records accessing pages by user size of 1.12 MB. The users access this Online Judge website to view their profile, to compile program's code, to contest their program, to ask advice program's issues and to solve their program problems. This web log record data includes the client's IP address, the request date and time, the requested page, the HTTP code and the status code accessed by the server. These include a detailed map of the client requesting the website from the web server.

3.2.1 Log File Format of Online Judge Web Log Data

```
10.128.2.1[29/Nov/2017:06:59:03 GET /home.php HTTP/1.1 200
```

- 10.128.2.1 - This is the IP address of the client requesting the server.
- [29/Nov/2017:06:59:03 (%t) -The time format resembles like [day/month/year: hour: minute: second zone]

- GET /home.php HTTP/1.1 - Requests sent from the client are given in double quotes. GET is the method used./home.php is the information requested by the client. The protocol used by the client is given as HTTP/1.1.
- 200 - It indicates that the code beginning with the status code sent by the server 200 indicates a successful response; 302 indicates a redirect; 401 indicates this error and 404 indicates a server error.

3.2.2 Preprocessing

Data preprocessing is pivotal in Web usage mining. This is a very complex process, accounting for 80% of the total mining process. Log files can be a source of noise, which can affect the results of mining operations. Irrelevant because it contains clear data, it is assumed that they need to be pre-processed. Before, applying any web mining algorithm, it is critical to filter and organize appropriate data. The goal of data preprocessing is to improve data quality and accuracy during mining. The preprocessing is divided into four stages: data cleaning, field extraction, user identification and session identification.

The information received from the online log file is unreliable, loud, and initially unsuitable for mining. To convert the data into an appropriate format for pattern recognition, preprocessing is necessary. Because the source of web logs data causes is combined with improper information, we begin in the pre-processing step by removing data, followed by data cleaning and data filtering. In Web usage mining, data pretreatment plays a key role. By applying web mining algorithms scheduled for the web server logs, it is utilized to sort and systematize only pertinent information. The cutting-edge server logs are organized into significant sessions before being used by WUM after being cleaned, formatted, and aggregated. Data cleansing, user identification, and session identification are the three sub steps that make up this stage.

(i) Removal of unsuccessful requests

The HTTP Status Code field is used to identify the status of the request when the server side did not successfully complete a user request or when the client attempted to issue a bad request. Table 3.1 shows the range of codes and their usages. Requests with irrelevant HTTP methods will be forcibly removed. Requests with GET or POST methods are important for analysis because they describe the actual user navigation behavior, so requests with other methods must be omitted. HTTP methods such as get,head,put,post,connect,trace,options, delete,propfind,cook, and so on are commonly found in log files. Some important methods are described below: GET: It is a vital method for retrieving documents from a Web server using a specific URI. It makes no other changes to the server's data. POST: A method for sending data to a Web server. It may update existing server data. HEAD: This method performs similar functionality to GET, except that the server only transfers headers and lines and not the entity body. PUT: This method is used to replace the target resource with the resource specified by the URI. DELETE: This method is used to delete the target resource specified by the URI. OPTIONS: A client uses this method to learn about the methods and options offered by a Web server. CONNECT: This method is used to connect to a specific Web server via HTTP Protocol.

(i) Removal of multimedia and other irrelevant files requests

Another important type of redundant requests occurs in log files. When a user requests a web page, multimedia objects embedded in the web page are automatically downloaded, even if the user did not explicitly request it. Because these requests do not reflect the actual behavior of the user, they should be deleted by checking the suffix of the Requested url field. Table 3.1 contains some irrelevant file extensions and their meanings, which extract log records from databases and remove requests for file extensions such as jpg, gif, and css. Another data cleaning algorithm is proposed, which reads records from the log table (stored in the database), removes records with URL suffixes such as jpg, gif, css, and records whose HTTP method is not GET, and the HTTP status code is other record of. over 200. For data cleaning, we use a combination of the two algorithms mentioned above, as discussed further below. The data cleaning algorithm receives input from the output of the data extraction algorithm, and removes irrelevant entries such as unsuccessful requests, irrelevant HTTP method requests, and requests for irrelevant files. After applying the data cleaning algorithm, the original log is reduced. Nearly 80%, which means that only 20% of the content in the raw log file is relevant.

File Type	Description
Jpeg,jpg,gif,png,tif,bmp	Image file
Mp3	Audio file
Swf	Flash animation file
Ico	Icon image file format
Cgi	Common gateway interface

Table 3.1 Description of irrelevant file extensions

3.2.3 Data Cleaning

In our proposed method, data cleaning is used to remove irrelevant transactions from log files that are useless for analysis to improve data quality. Data cleansing is generally a site-specific process where entries are removed based on analysis type, i.e. we do not remove image file requests when performing analysis on image-specific sites. In our case, we considered unsuccessful requests, requests containing multimedia and other irrelevant files, and HTTP methods other than GET or POST.

```
Input : Web Server Log File
Output: Cleaned Log Record
Step 1 : Read web log record from Web Server Log File
Step 2: IF ((LogRecord.status <>(200,302)
            AND ((LogRecord.method =‘GET’))
            AND ((LogRecord.stem  (php,css,js,txt)))
             THEN Insert LogRecord into LogDatabase
             End of IF Condition
Step 3: Repeat the above two steps until endof  (Web
             Server Log File)
Step 4 : Stop the process.
```

Table 3.2 Algorithm for Data Cleaning

3.2.4 User Identification

User identification is the process of identifying specific users by using their IP addresses. To identify unique users, the following rule apply:

1) If the IP address changes, there is a new user;

```
Input   : Cleaned Log Record
Output : Unique Users Database
Step 1 : Initialize  IPList =0; UsersList = 0; No-of-users = 0;
Step 2 : Read Record from LogDatabase
Step 3 : IF Record.IP address is not in IPList THEN
             Add new Record.Ip address into IPList
             increment count of No-of-users
               insert new user  into UserList
               End of IF
Step 4 : Repeat the above steps 2 to 3 until endof(Log Database)
Step 5 : Stop the process.
```

Table 3.3 Algorithm for User Identification

The goal of the user identification process is to find the various users in the web access log file. Different users are distinguished by their Internet Protocol (IP) addresses. The identification of the user is also critical because it has a large impact on the quality of the pattern discovery result.

3.2.5 Page Identification

The page identification step identifies individual page by using their URL name. If there is new URL, there is new page.

```
Input   : Cleaned Log Record
Output : Unique Pages Database
Step 1 : Initialize  PageList =0; No-of-Page=0;
Step 2 : Read Record from LogDatabase
Step 3 : IF Record.URL name (Web Page) is not in PageList
          THEN
            Assign Page Id for new Web Page
          Insert Page id into PageList
          increment count of No-of-Page
Step 4 : Repeat the above steps 2 to 3 until endof(Log Database)
Step 5 : Stop the process.
```

Table 3.4 Algorithm for Page Identification

3.2.6 Session Identification

The term "session" refers to the amount of time a person spends on a website. Each user's page access is divided into several sessions using session identifier. For the same users, many sessions are conceivable. For locating or creating the user's sessions, which depend on both time and navigation, there are two options. In this study, session identifiers are generated by calculating the difference between two time stamps belonging to the same user.

Throughout web servers, user clicks are often referred to as click streams, and a set is defined as a user machine. A category in which web pages are viewed by a single user.

An even and static delay is the foundation of traditional session identification techniques. While the gap between two queries is longer than the timeout, the new session is unwavering.

```
Session Identification Algorithm

Input    : User Identification Database

Output : Session Database

Step 1 : Initialize SessionList = 0; No-of-Sessions = 0;

Step 2 : Read LogRecord from  User Database

Step 3 : IF  LogRecord.UserID.equal(NewLogRecord.UserID)

         AND (NewLogRecord.timetaken-LogRecord.timetaken>5))

         THEN

         Get Url address of corresponding Session and

         Insert into SessionList

         increment No-of-Sessions

         End of IF

Step 4 : Repeat the above steps 2 and 3 untill endof (Logrecord)

Step 5 : End of process.
```

Table 3.5 Algorithm for Session Identification

3.3 Pattern Discovery in Web Usage Mining

Pattern Discovery Phase, choosing necessary user patterns should be done after every preprocessing stage. However, this user pattern selection process will be acceptable the extracted web access log files after cleaning process. The extraction process also followed certain rules for categorizing the cleaned log files. This procedure identifies required sessions and users. Each session is separated by a five-minute interval based on the user's visit to the web page. The process of discovering patterns is also known as rule mining. This process requires many rules and classifies web user paths based on the rules specifically mentioned in the user behavior extraction process. For grouping these paths, web mining researchers employ a variety of algorithmic techniques. The paths that have been grouped or classified are used in the analyzing stage.

After the data preprocessing phase, need to apply the pattern detection method. This phase consists of different techniques derived from different disciplines such as statistics, machine learning, data mining, and pattern recognition, and applies to web domains and available data. The task for discovering patterns provides several techniques such as statistical analysis, association rules, sequential pattern analysis, and clustering. Some techniques make it easier to discover patterns in the processed data.

After pre-processing the data from the server log file, the page access links are got by visitors. If the page links are from the same IP address: although the system can determine if the user is the same person, the number of visitors to the page will increase. If the page link is from a different IP address: The system will increase the number of visitors. And then the user’s behavior or access page links will be found from these transaction lists using one of the frequent itemset mining algorithm. A lot of algorithms have already been designed for generating frequent itemset. One of the Eclat Algorithm have been used.

3.4 Eclat Algorithm in Association Mining Rule

In the majority of organizations, frequent itemset calculating is necessary. Items that appear frequently in the database are known as frequent itemset. For creating frequent itemset, many methods have already been developed. Eclat algorithm is one of the algorithms have been employed. When searching for entries in a database, the Eclat algorithm employs a vertical dataset and a bottom-up methodology.

A very simple algorithm is Apriori. However, calculations of frequently occurring itemset take a long time. We must compute support and confidence in this algorithm and the database must be scanned the database. Thus, the FP growth algorithm was created. Frequent pattern growth is referred to as FP growth. Database scanning is necessary twice in this technique. Time usage has increased. Eclat algorithm was created to get rid of these restrictions. Some association rule mining algorithms use a horizontal data format and some use a vertical data format to generate frequent itemset. Eclat cannot use horizontal databases. This vertical approach to the ECLAT algorithm makes it a faster algorithm than Apriori.In Apriori algorithm, database needed to scan database again and again for finding frequent itemset, this limitation is reduced by using vertical dataset in Eclat.

Eclat needs to scan the database only once. All the data is stored in vertical form. Bottom up approach is used for searching items in the database. Searching starts from bottom to top. Only support is calculated in this algorithm. But it takes more time than top down approach. Firstly, all items are calculated individually support and support can be decided by user. After calculating support of all items, support of items will be compared with minimum decided support. The items which has support more than or equal to decided support are frequent itemset. And the following table 3.6 ECLAT algorithm is used step by step by paring items to generate frequent itemset. This is the whole process of Eclat algorithm.

```
Input: $D_k$ = {I1, I2, ..., In} // cluster of frequent k-itemsets.
 Output: Frequent l-itemsets.
Bottom-Up ($D_k$) {
1. for all $I_i$ € $D_k$
2. $D_{k+1} = \emptyset$;
3. for all $I_j$ € $D_k$ , $i < j$
4. $N = Ii \cap Ij$;
5. if N .sup>= min_sup then
6. $D_{k+1} = D_{k+1}$ U N;
7. end
8. end
9. end
10. if $D_{k+1} \neq \emptyset$ then
11. Bottom-Up ($D_{k+1}$);
12. end
13.}
```

Table 3.6 Explanation of ECLAT Algorithm

Take D_k as input. Output will be Set of frequent itemsets. In Bottom-Up (D_k), for loop is defined in which all items belong to database D_k under first step. In step2, take D_{k+1} as empty database. In step3, check that item exists in the database D_k or not. If item exists in the database D_k then calculate support of item in step4 where support means how many times item occurs in database D_k. In step5, compare support of items individually with minimum decided support. In step6, put those items whose support is more than then minimum support in database D_{k+1}. In step7, check that database D_{k+1} is empty or not. If it is not empty then start the same procedure for another database.

3.4.1 Advantages of ECLAT Algorithm

1. The Eclat algorithm uses the Depth-First Search method, which has lower memory needs than Apriori.
2. The Eclat algorithm is typically faster than the Apriori approach since it does not constantly scan the data to find frequent itemsets.
3. As long as the dataset is not too large, the Eclat method beats the Apriori algorithm.
4. Eclat algorithm scans only the currently generated dataset that is scanned in the Eclat algorithm.

3.5 Calculation of ECLAT ALGORITHM

Transaction	ITEM 1	ITEM 2	ITEM 3	ITEM 4	ITEM 5
T01	0	0	1	0	0
T02	0	1	1	1	0
T03	1	1	0	1	1
T04	1	1	0	1	1
T05	1	0	0	1	0
T06	0	1	1	1	1
T07	1	0	1	0	0
T08	0	0	1	1	1

Table 3.7 Original Database

ITEMS	Transaction
ITEM 1	{ T03, T04, T05, T07 }
ITEM 2	{ T02,T03,T04,T06 }
ITEM 3	{ T01,T02,T06,T07,T08 }
ITEM 4	{ T02,T03, T04, T05, T06, T08 }
ITEM 5	{ T03, T04, T06, T08 }

Table 3.8 Vertical Format with Items

Itemset	Support
ITEM 1	4
ITEM 2	4
ITEM 3	5
ITEM 4	6
ITEM 5	4

Table 3.9 Calculating Support

Average support is 3. Take items which has support equal to or more than 3.

Itemset	Support
ITEM 1, ITEM 2	2
ITEM 1, ITEM 3	1
ITEM 1, ITEM 4	1
ITEM 1, ITEM 5	2
ITEM 2, ITEM 3	2
ITEM 2, ITEM 4	4
ITEM 2, ITEM 5	3
ITEM 3, ITEM 4	3
ITEM 3, ITEM 5	2
ITEM 4, ITEM 5	4

Table 3.10 After Comparing Support

Itemset	Support
ITEM 2, ITEM 4	4
ITEM 2, ITEM 5	3
ITEM 3, ITEM 4	3
ITEM 4, ITEM 5	4

Table 3.11 Paired itemset after checking support

Itemset	Support
ITEM 2, ITEM 4, ITEM 5	3
ITEM 3, ITEM 4, ITEM 5	2

Table 3.12 Paired three itemset

Itemset	Support
ITEM 2, ITEM 4, ITEM 5	3

Table 3.13 The final result of frequent itemset

In table 3.7, 8 transactions are taken. 5 items are sold in 101 to 108 transactions. 0 represent item does not exist in the database and 1 represent item exists in the database. Table 3.8 shows support of each item in every transaction i.e. items come in how many transactions and compare support of each item with decided support. Table 3.9 shows only those items whose support is equal to and greater than decided support. In table 3.10, items of table 3.9 have been paired and calculated their support i.e. in total numbers of transactions in which their pairs come collectively. Table 3.11 shows only those paired items whose support is equal to and greater than decided support. In table 3.12, items of table 3.11 have been paired and calculated their support i.e in total numbers of transactions in which their pairs come collectively. Table 3.13 shows only those paired items whose support is equal to and greater than decided support.

CHAPTER 4
DESIGN AND IMPLEMENTATION OF SYSTEM

This chapter represents overview design of the system, data pre-processing, step by step calculation of ECLAT Algorithm on web log data, user interface design. This chapter begins with an overview of the system design with diagrams and detailed explanations. Second, the data preprocessing procedures for the web log data and clean datasets are described. Thirdly, it calculates frequent patterns on web log data using association rule algorithm. Finally, the user interface design of the system is presented with step-by-step detailed explanatory diagrams.

4.1 Overview Design of the System

In overview design of the system (Figure 4.1), the main process is to generate frequent pattern means that association rules. Firstly, the user input web log file then the system checks log file and removes unnecessary data as data preprocessing algorithms. After that, preprocessing can be done with data cleaning, user identification, page identification and session identification. Preprocessing phase is the most important one because it makes the data with good quality. Then, the data from processing phase is stored into the database. Secondly, the system stores clean data to database and selects data to change the vertical data format and to generate frequent patterns using ECLAT Algorithm. Frequent pattern mining also used to find information like set of pages repeatedly accessed together by users. Thirdly, the result from mining frequent pattern is measured performance which means measures the confidence on frequent pattern. Finally, it generates the output that is strong association rules.

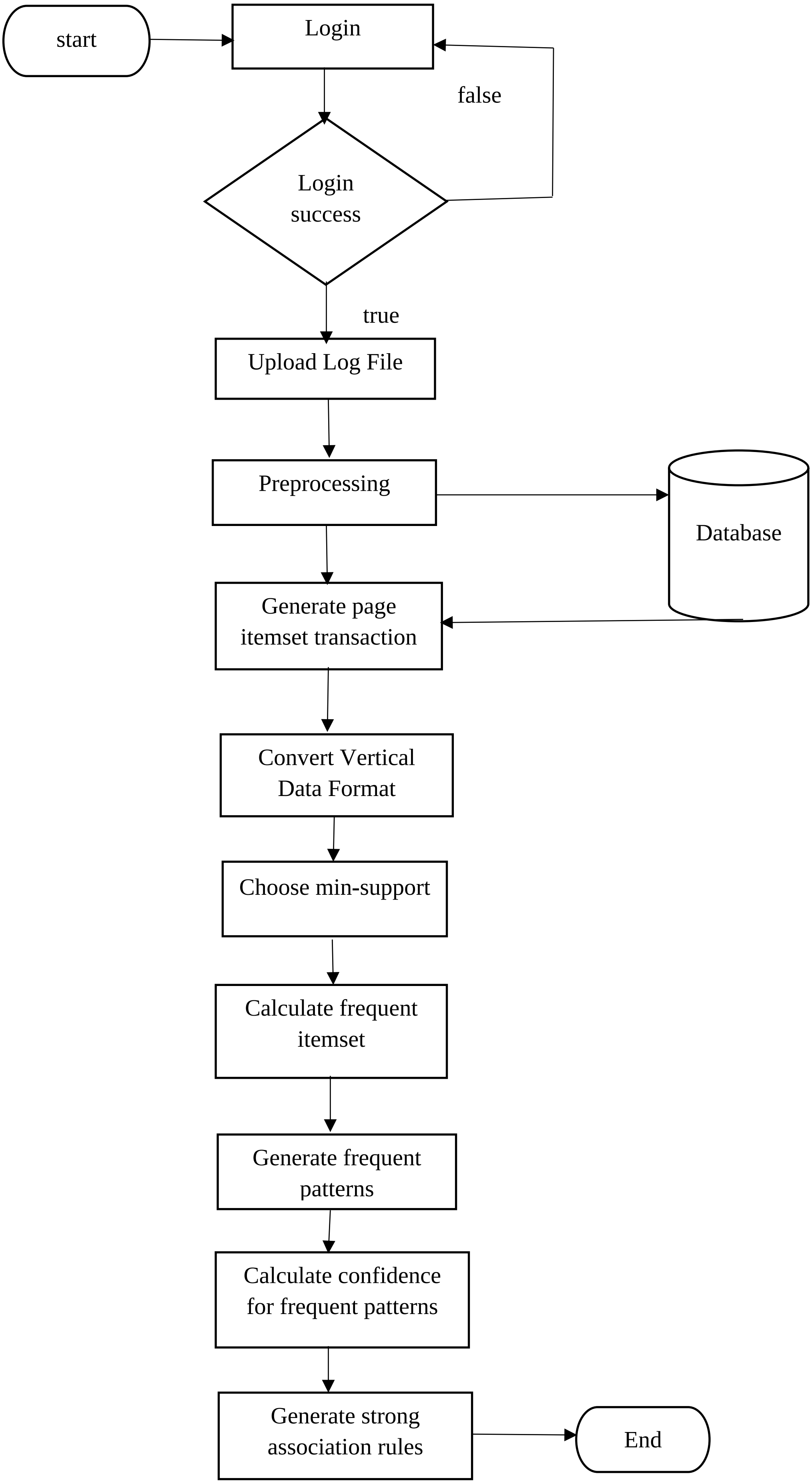

start
Login
false
Login success
true
Upload Log File
Preprocessing
Database
Generate page itemset transaction
Convert Vertical Data Format
Choose min-support
Calculate frequent itemset
Generate frequent patterns
Calculate confidence for frequent patterns
Generate strong association rules
End

Figure 4.1 Overview Design of the System

4.2 Preprocessing

Data preprocessing includes removing useless requests from log files. Since all log entries are invalid, the irrelevant entries need to be eliminated. Typically, this process removes requests for non-analytical resources, such as images, multimedia files, and page style files. For example, requests for graphical page content (*.jpg and *.gif images) and requests for any other files that might be included in a web page, or even navigation sessions performed by bots and web spiders. By filtering out useless data, the log file size can be reduced to use less storage space and facilitate upcoming tasks. By cleaning the data, we can create a database according to the application, including user identification, session identification, page identification, frequent patterns and other information.

```
10.128.2.1[29/Nov/2017:06:58:55GET /login.php HTTP/1.1200
10.128.2.1[29/Nov/2017:06:59:02POST /process.php HTTP/1.1302
10.128.2.1[29/Nov/2017:06:59:03GET /home.php HTTP/1.1200
10.131.2.1[29/Nov/2017:06:59:04GET /js/vendor/moment.min.js HTTP/1.1200
10.130.2.1[29/Nov/2017:06:59:06GET /bootstrap-3.3.7/js/bootstrap.js HTTP/1.1200
10.130.2.1[29/Nov/2017:06:59:19GET/profile.php?user=bala HTTP/1.1200
10.128.2.1[29/Nov/2017:06:59:19GET /js/jquery.min.js HTTP/1.1200
10.131.2.1 [29/Nov/2017:06:59:19GET /js/chart.min.js HTTP/1.1   200
10.131.2.1 [29/Nov/2017:06:59:30GET /edit.php/ HTTP/1.1200
10.130.2.1 [29/Nov/2017:15:04:41GET /home.php HTTP/1.200
10.130.2.1 [29/Nov/2017:15:05:11GET /contest.php HTTP/1.1200
10.131.2.1[29/Nov/2017:15:05:30GET /contest.php HTTP/1.1200
```

Table 4.1 Raw data of the web log file

4.2.1 Data Cleaning

Data cleaning means remove the irrelevant information from the original Web log file and restore the Web log as database which is suitable for user identification, session identification and path complement.

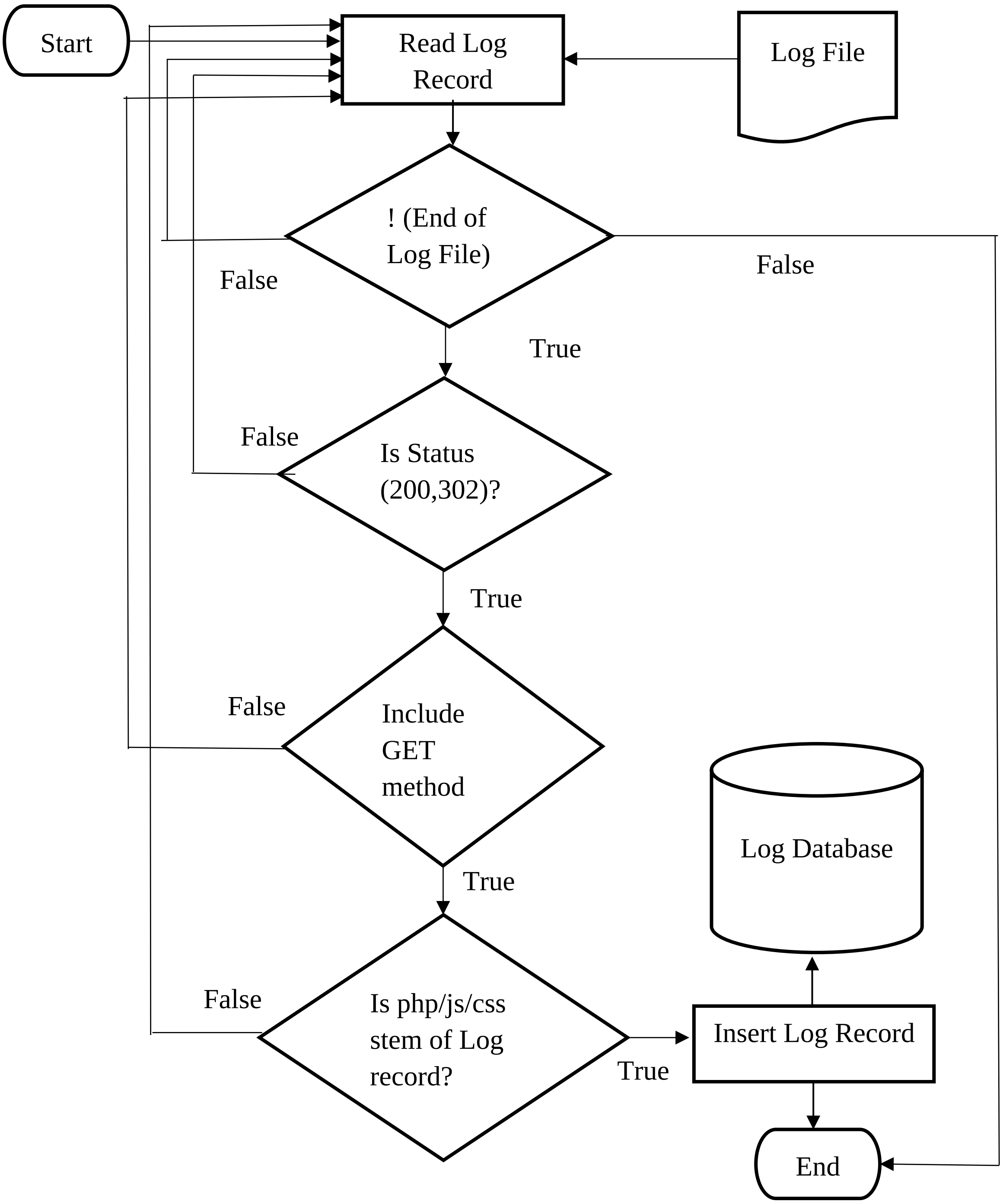


Figure 4.2 Data Flow Diagram of the Data Cleaning

According to this proposed algorithm, it will remove irrelevant data. This algorithm will remove the accessorial entries like jpg, gif, png files and entries with status code is not 200 and 302.

IP Address	Time	Web Pages
10.128.	29/Nov/2017:06:58:55	login.php
10.128.2.1	29/Nov/2017:06:59:03	home.php
10.128.2.1	29/Nov/2017:06:59:04	js/vendor/moment.min.js
10.130.2.1	29/Nov/2017:06:59:06	bootstrap-3.3.7/js/bootstrap.js
10.130.2.1	29/Nov/2017:06:59:19	profile.php
10.128.2.1	29/Nov/2017:06:59:19	js/jquery.min.js
10.131.2.1	29/Nov/2017:06:59:19	js/chart.min.js
---	---	--
10.131.2.1	29/Nov/2017:06:59:30	edit.php
10.131.2.1	29/Nov/2017:06:59:37	login.php

Table 4.2 Result of Data Cleaning on Web Log Data

4.2.2 User Identification

User identification are intended to get into each user's access credentials; it then creates a group of users and provides them with a personalized service. Each user has a unique IP address, and each IP address represents a user.

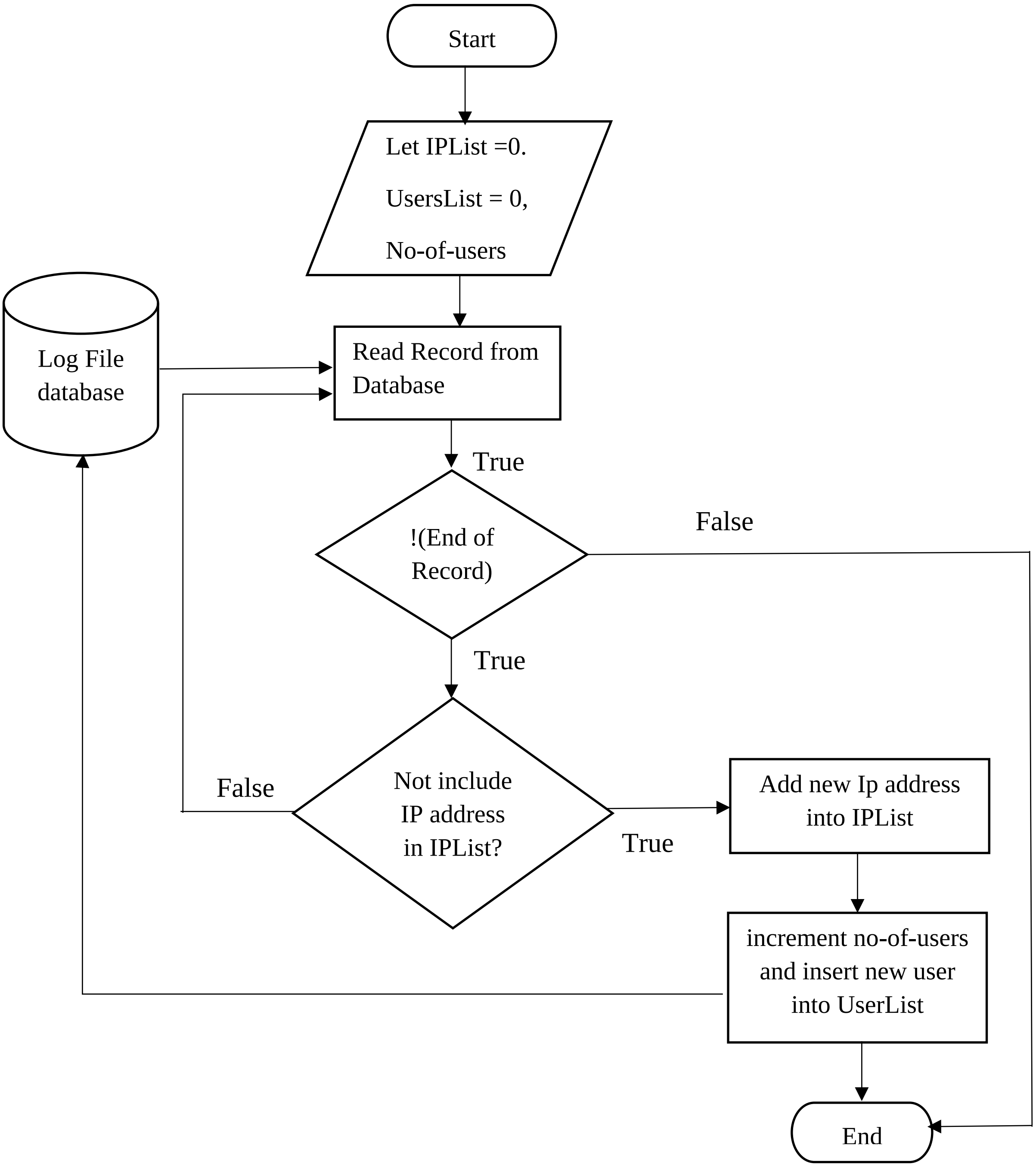

Start
Let IPList =0.
UsersList = 0,
No-of-users
Log File database
Read Record from Database
True
!(End of Record)
False
True
Not include IP address in IPList?
False
True
Add new Ip address into IPList
increment no-of-users and insert new user into UserList
End

Figure 4.3 Data Flow Diagram of User Identification

User 1 – 10.128.2.1		
IP Address	**Date and Time**	**Web Pages**
10.128.2.1	29/Nov/2017:06:58:55	login.php
10.128.2.1	29/Nov/2017:06:59:03	home.php
10.128.2.1	29/Nov/2017:06:59:19	js/jquery.min.js
10.128.2.1	29/Nov/2017:13:38:20	css/style.css
10.128.2.1	29/Nov/2017:13:38:20	css/font-awesome.min.css
10.128.2.1	29/Nov/2017:13:38:21	bootstrap-3.3.7/js/bootstrap.min.js
-----------	--------------------------	-------------------------
10.128.2.1	29/Nov/2017:13:49:37	compiler.php

Table 4.3 Result of the User Identification

4.2.3 Page Identification

The page identification step uses the URL name to identify individual pages. There is a new page if the URL changes. In this proposed system, each URL name will be defined as a new page. The algorithm will read each cleaned log record file, as shown in the flow chart. When a new URL name is introduced, it defines the symbol for each page.

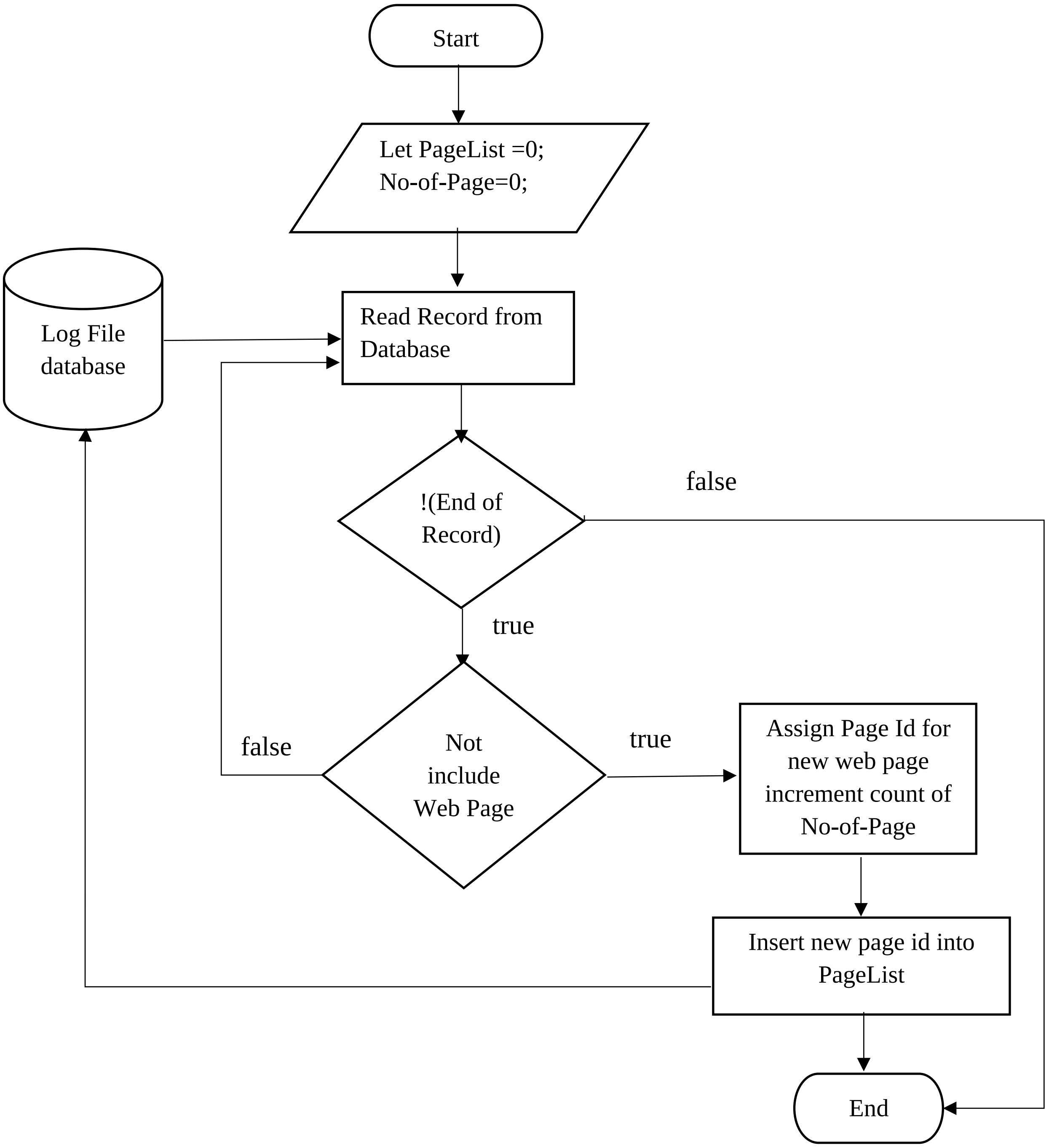

Start
Let PageList =0;
No-of-Page=0;
Log File database
Read Record from Database
!(End of Record)
false
true
Not include Web Page
false
true
Assign Page Id for new web page increment count of No-of-Page
Insert new page id into PageList
End

Figure 4.4 Data Flow Diagram of Page Identification

Page Symbol	Web Pages
P01	/login.php
P02	/home.php
P03	/js/vendor/moment.min.js
P04	/bootstrap-3.3.7/js/bootstrap.js
P05	/profile.php
--------------------------------	--------------------------------------
P20	/countdown.php
P21	/compiler.php
P22	/details.php
P23	/contest.php

Table 4.4 Result of Page Identification

4.2.4 Session Identification

The proposed user session Identification algorithm is given below. A user session can be characterized as a collection of pages that the same person visits during a single visit to a website. The session identification method is as follows. (1) If a new user is present, there is a new session. (2) If the reference page for a given user session is null, it is assumed that a new session has begun. (3) If the duration between page requests is longer than a predetermined threshold. According to this algorithm, it will define new session that the page request exceeds over 5 min.

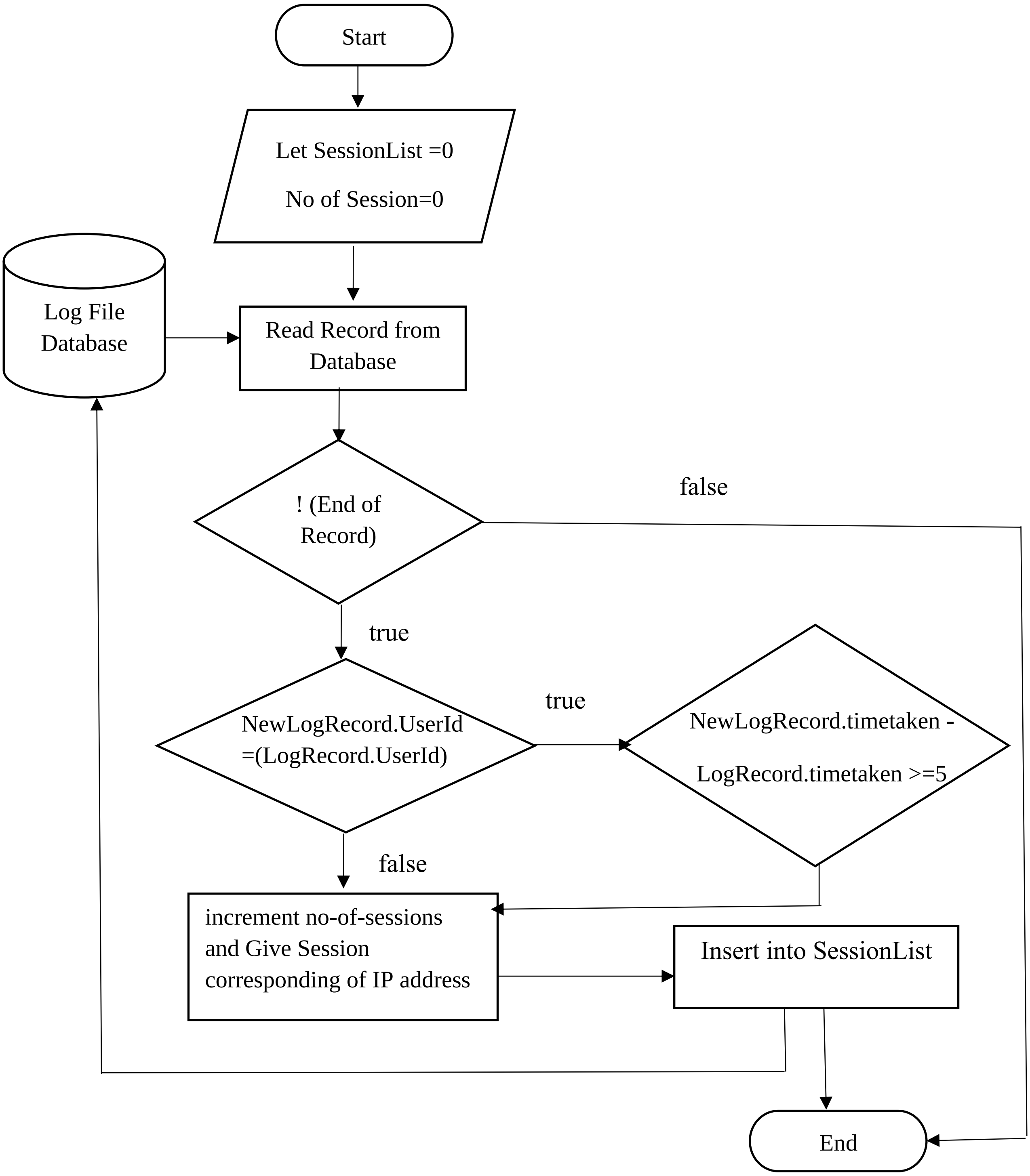

Start
Let SessionList =0
No of Session=0
Log File Database
Read Record from Database
! (End of Record)
false
true
NewLogRecord.UserId =(LogRecord.UserId)
true
NewLogRecord.timetaken - LogRecord.timetaken >=5
false
increment no-of-sessions and Give Session corresponding of IP address
Insert into SessionList
End

Figure 4.5 Data Flow Diagram of Session Identification

User 1 – Session 1 (p01,p02,p09,p06)		
Ip Address	Date	Web Pages
10.128.2.1	29/Nov/2017 06:58:55	/login.php
10.128.2.1	29/Nov/2017 06:59:03	/home.php
10.128.2.1	29/Nov/2017 06:59:03	/contestproblem.php
10.128.2.1	29/Nov/2017 06:59:19	/js/jquery.min.js

Table 4.5 Result of Session Identification of each user

4.3 Implementation of the System

The system is implemented with the node.js as shown in Figure 4.6. In this page, the user can upload raw data of web log file from browse. Firstly, the user will choose log file by clicking browse button and then the user can upload this log file by pressing upload button.

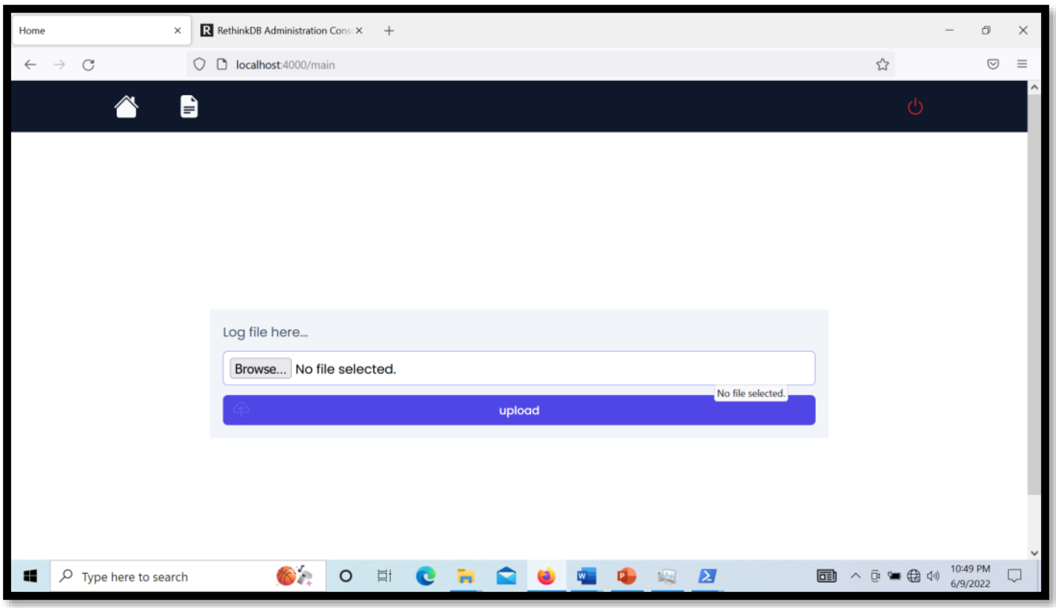
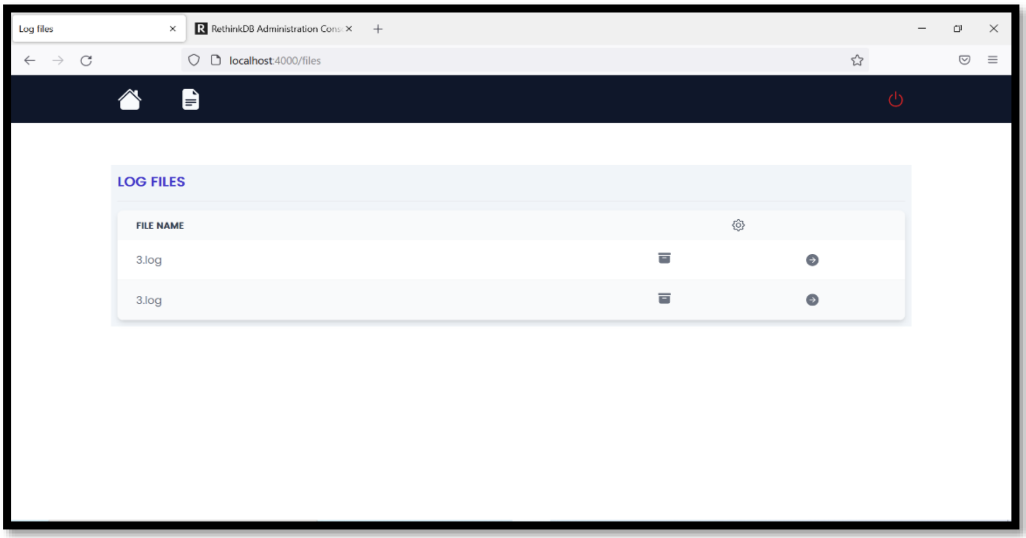


Figure 4.6 Home Page of the system

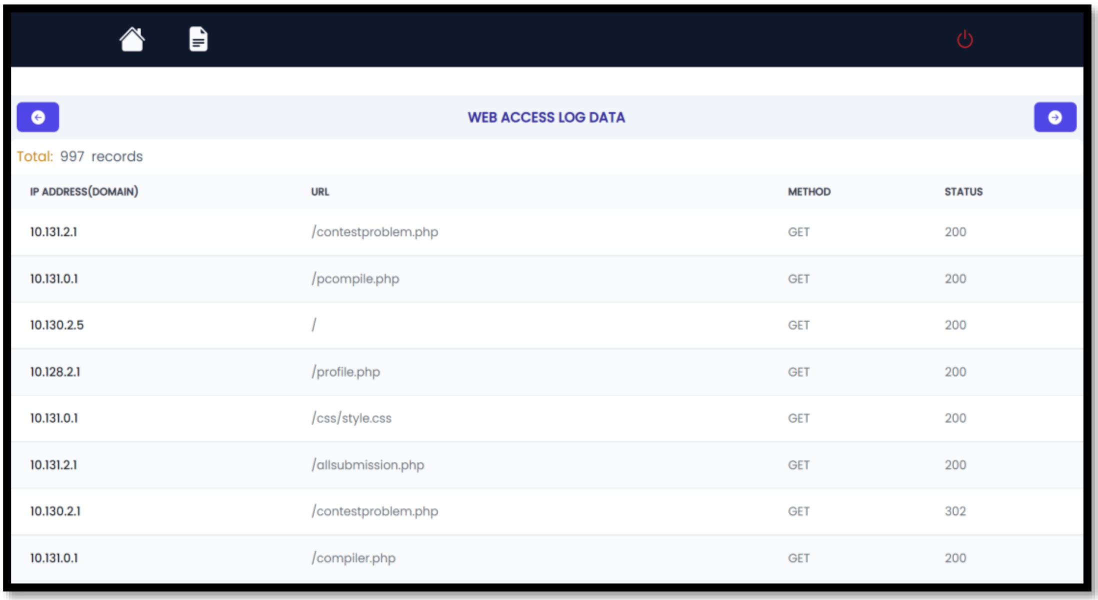


LOG FILES

FILE NAME		
3.log		
3.log		

Figure 4.7 File Upload of the system

After uploading log file, the user can view log file page as shown in figure 4.7. In this page, the system will display the user loaded log files so the user can view log files or can delete these files by clicking delete button. After that the user can view data from log file by clicking arrow button.



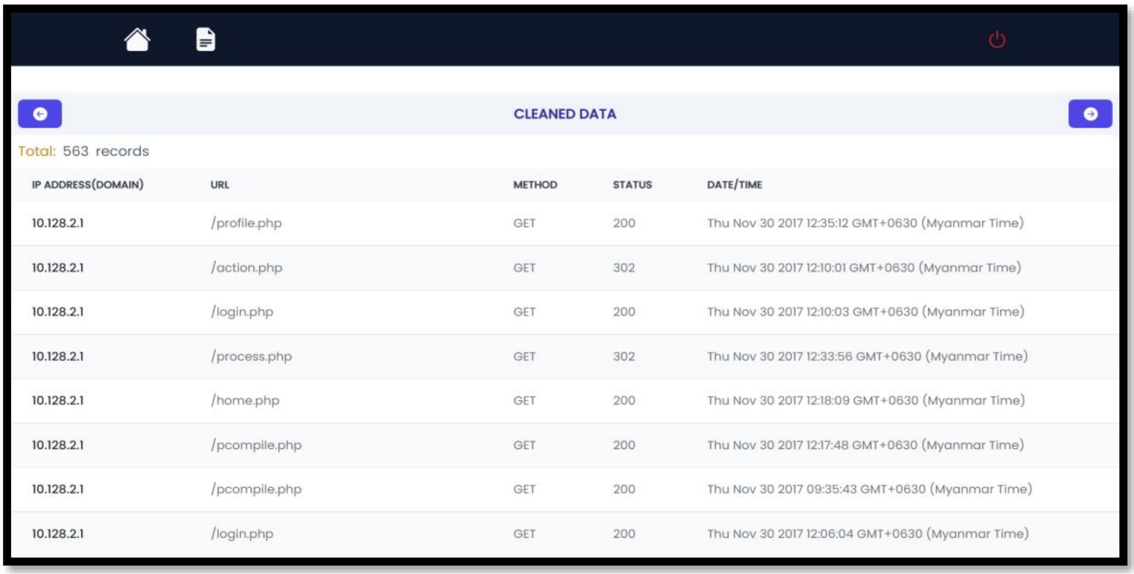
WEB ACCESS LOG DATA

Total: 997 records

IP ADDRESS(DOMAIN)	URL	METHOD	STATUS
10.131.2.1	/contestproblem.php	GET	200
10.131.0.1	/pcompile.php	GET	200
10.130.2.5	/	GET	200
10.128.2.1	/profile.php	GET	200
10.131.0.1	/css/style.css	GET	200
10.131.2.1	/allsubmission.php	GET	200
10.130.2.1	/contestproblem.php	GET	302
10.131.0.1	/compiler.php	GET	200

Figure 4.8 Displaying Web Log Record

In this page, the system represents log data each title with columns. There are four titles such as IP address, page URL, method and status code.



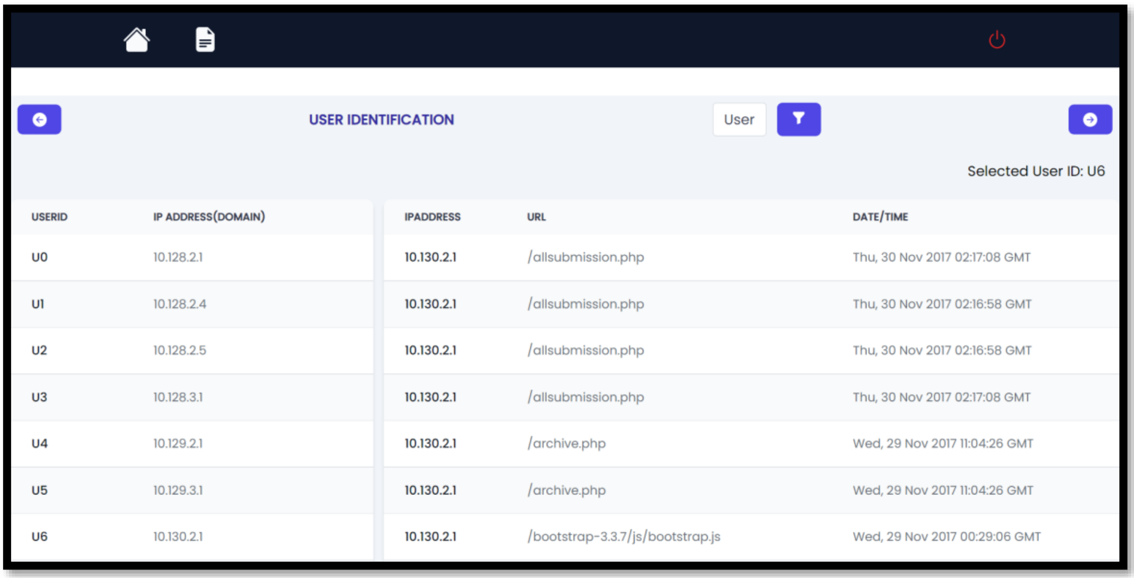
CLEANED DATA

Total: 563 records

IP ADDRESS(DOMAIN)	URL	METHOD	STATUS	DATE/TIME
10.128.2.1	/profile.php	GET	200	Thu Nov 30 2017 12:35:12 GMT+0630 (Myanmar Time)
10.128.2.1	/action.php	GET	302	Thu Nov 30 2017 12:10:01 GMT+0630 (Myanmar Time)
10.128.2.1	/login.php	GET	200	Thu Nov 30 2017 12:10:03 GMT+0630 (Myanmar Time)
10.128.2.1	/process.php	GET	302	Thu Nov 30 2017 12:33:56 GMT+0630 (Myanmar Time)
10.128.2.1	/home.php	GET	200	Thu Nov 30 2017 12:18:09 GMT+0630 (Myanmar Time)
10.128.2.1	/pcompile.php	GET	200	Thu Nov 30 2017 12:17:48 GMT+0630 (Myanmar Time)
10.128.2.1	/pcompile.php	GET	200	Thu Nov 30 2017 09:35:43 GMT+0630 (Myanmar Time)
10.128.2.1	/login.php	GET	200	Thu Nov 30 2017 12:06:04 GMT+0630 (Myanmar Time)

Figure 4.9 Viewing Cleaned Log Data

In this page, the user can view cleaned data records. The system removes irrelevant data that the records are not GET method and status code 200.So, these records are completely cleaned data records. The user can also view each user (IP address) accessed page URL record together with datetime.



USER IDENTIFICATION

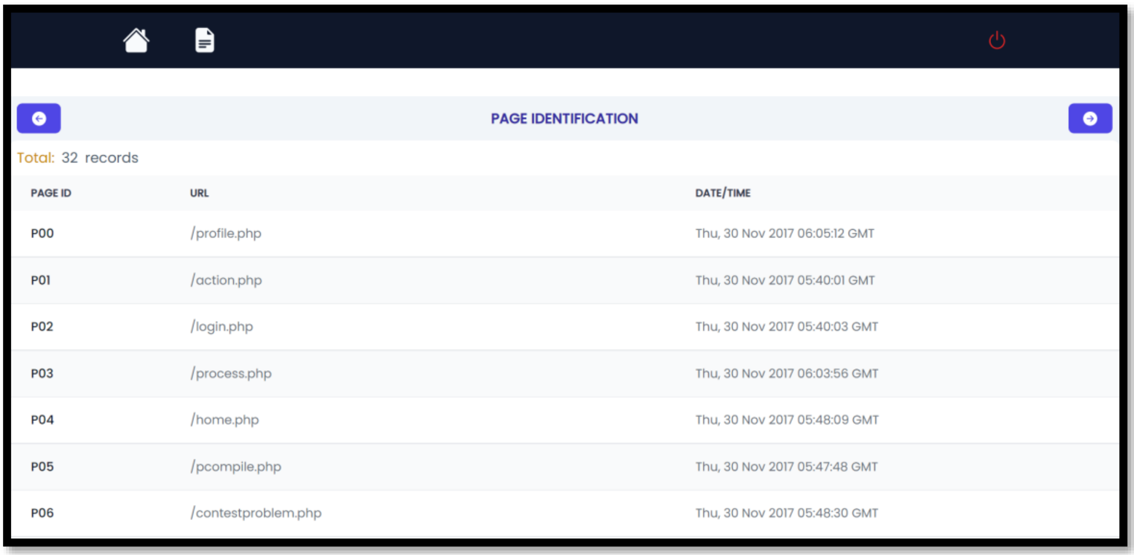
Selected User ID: U6

USERID	IP ADDRESS(DOMAIN)
U0	10.128.2.1
U1	10.128.2.4
U2	10.128.2.5
U3	10.128.3.1
U4	10.129.2.1
U5	10.129.3.1
U6	10.130.2.1

IPADDRESS	URL	DATE/TIME
10.130.2.1	/allsubmission.php	Thu, 30 Nov 2017 02:17:08 GMT
10.130.2.1	/allsubmission.php	Thu, 30 Nov 2017 02:16:58 GMT
10.130.2.1	/allsubmission.php	Thu, 30 Nov 2017 02:16:58 GMT
10.130.2.1	/allsubmission.php	Thu, 30 Nov 2017 02:17:08 GMT
10.130.2.1	/archive.php	Wed, 29 Nov 2017 11:04:26 GMT
10.130.2.1	/archive.php	Wed, 29 Nov 2017 11:04:26 GMT
10.130.2.1	/bootstrap-3.3.7/js/bootstrap.js	Wed, 29 Nov 2017 00:29:06 GMT

Figure 4.10 User Identification Page

In this page the system defines that each IP address is one user. The end user can filter each user ‘s accessed page URL records. When the end user searches each user's accessed record, he or she select one user's id. So, this page is called user identification page.



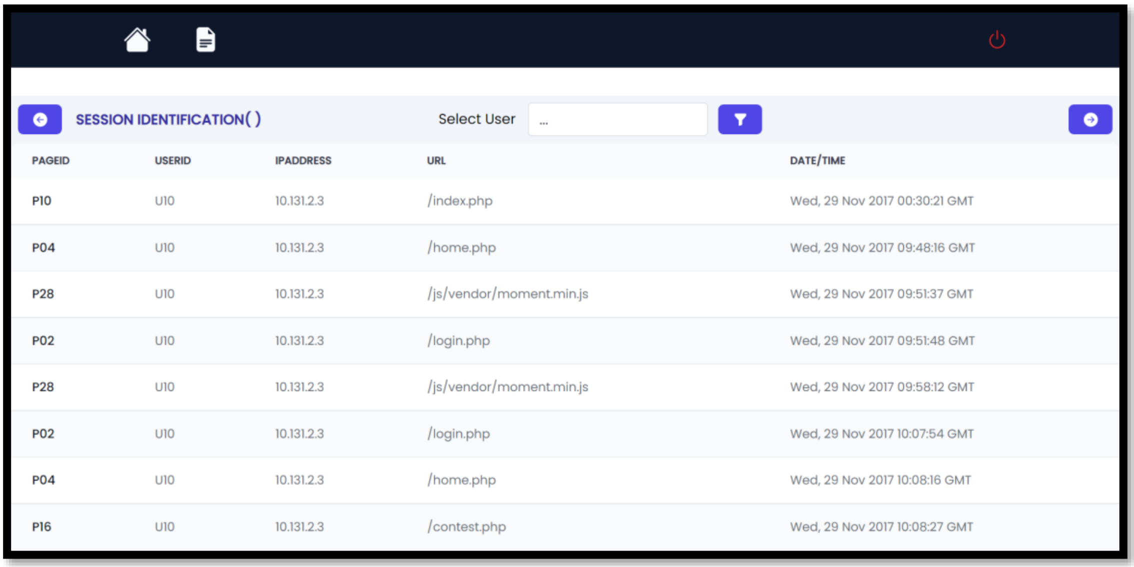
PAGE IDENTIFICATION

Total: 32 records

PAGE ID	URL	DATE/TIME
P00	/profile.php	Thu, 30 Nov 2017 06:05:12 GMT
P01	/action.php	Thu, 30 Nov 2017 05:40:01 GMT
P02	/login.php	Thu, 30 Nov 2017 05:40:03 GMT
P03	/process.php	Thu, 30 Nov 2017 06:03:56 GMT
P04	/home.php	Thu, 30 Nov 2017 05:48:09 GMT
P05	/pcompile.php	Thu, 30 Nov 2017 05:47:48 GMT
P06	/contestproblem.php	Thu, 30 Nov 2017 05:48:30 GMT

Figure 4.11 Displaying Page Identification

In this page identification, the system defines the symbol of each page as page ID. The system displays all page URL in the web log file. This page is also called page identification.



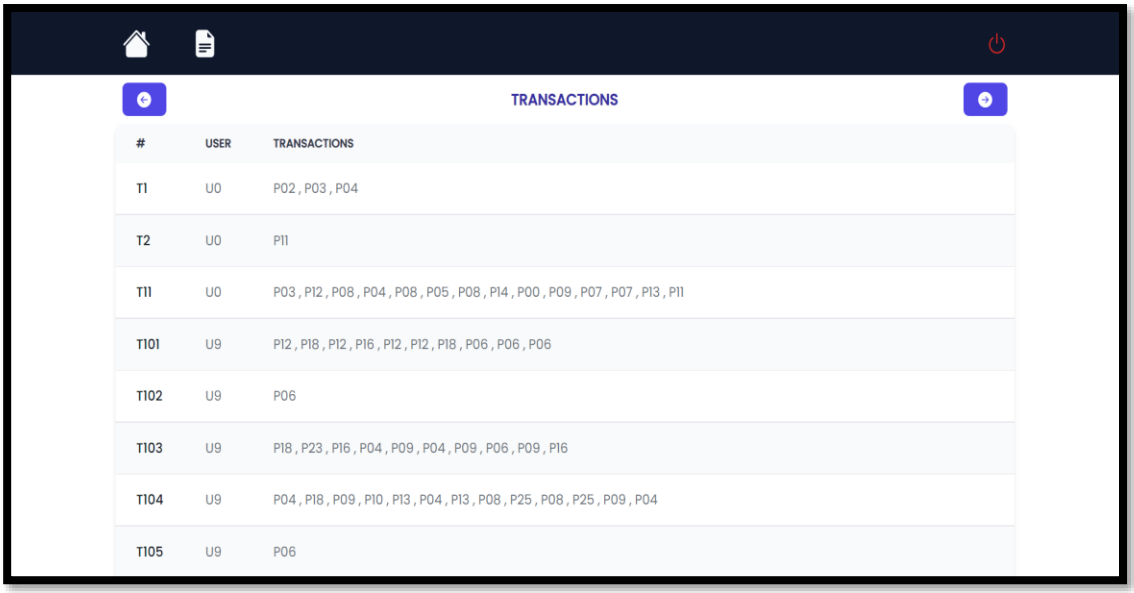
SESSION IDENTIFICATION( )

Select User ...

PAGEID	USERID	IPADDRESS	URL	DATE/TIME
P10	U10	10.131.2.3	/index.php	Wed, 29 Nov 2017 00:30:21 GMT
P04	U10	10.131.2.3	/home.php	Wed, 29 Nov 2017 09:48:16 GMT
P28	U10	10.131.2.3	/js/vendor/moment.min.js	Wed, 29 Nov 2017 09:51:37 GMT
P02	U10	10.131.2.3	/login.php	Wed, 29 Nov 2017 09:51:48 GMT
P28	U10	10.131.2.3	/js/vendor/moment.min.js	Wed, 29 Nov 2017 09:58:12 GMT
P02	U10	10.131.2.3	/login.php	Wed, 29 Nov 2017 10:07:54 GMT
P04	U10	10.131.2.3	/home.php	Wed, 29 Nov 2017 10:08:16 GMT
P16	U10	10.131.2.3	/contest.php	Wed, 29 Nov 2017 10:08:27 GMT

Figure 4.12 Displaying Session Identification

In this session identification page, the system clearly displays each user's visited page url with datetime. The user changes his page access time so the system divides user's accessed time as session. The system defines that the new session if the new user access or the same user the duration of no accessing time is over five minutes. The end user can also search each user's session with information such as page id, user id, page url and date/time.

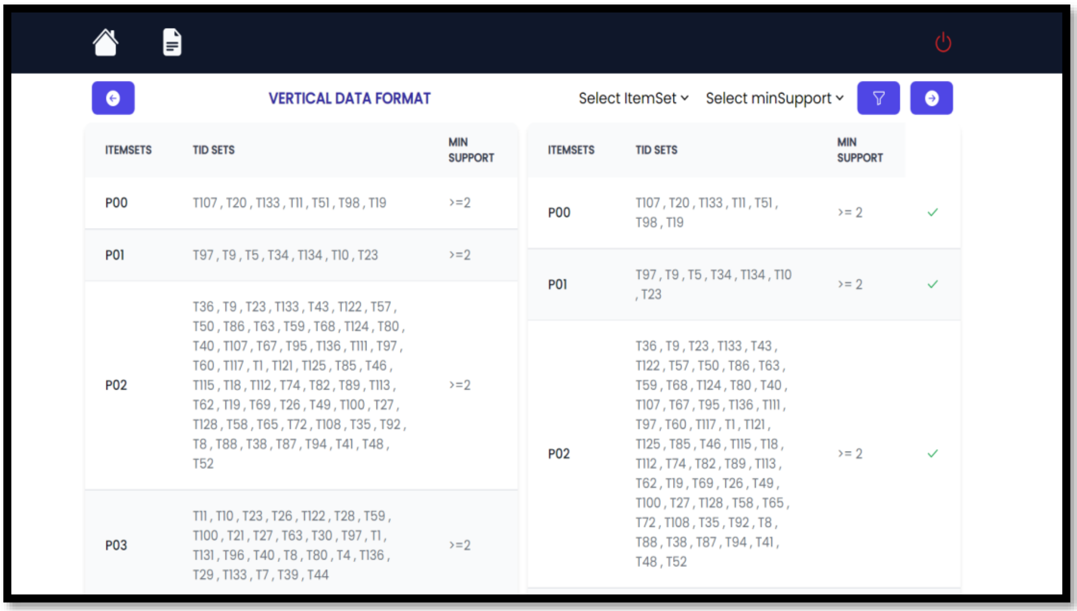


TRANSACTIONS

#	USER	TRANSACTIONS
T1	U0	P02 , P03 , P04
T2	U0	P11
T11	U0	P03 , P12 , P08 , P04 , P08 , P05 , P08 , P14 , P00 , P09 , P07 , P07 , P13 , P11
T101	U9	P12 , P18 , P12 , P16 , P12 , P12 , P18 , P06 , P06 , P06
T102	U9	P06
T103	U9	P18 , P23 , P16 , P04 , P09 , P04 , P09 , P06 , P09 , P16
T104	U9	P04 , P18 , P09 , P10 , P13 , P04 , P13 , P08 , P25 , P08 , P25 , P09 , P04
T105	U9	P06

Figure 4.13 Displaying Page Transaction List of each user

In this page, the end user can view each user's accessed page url lists as transaction of page url.



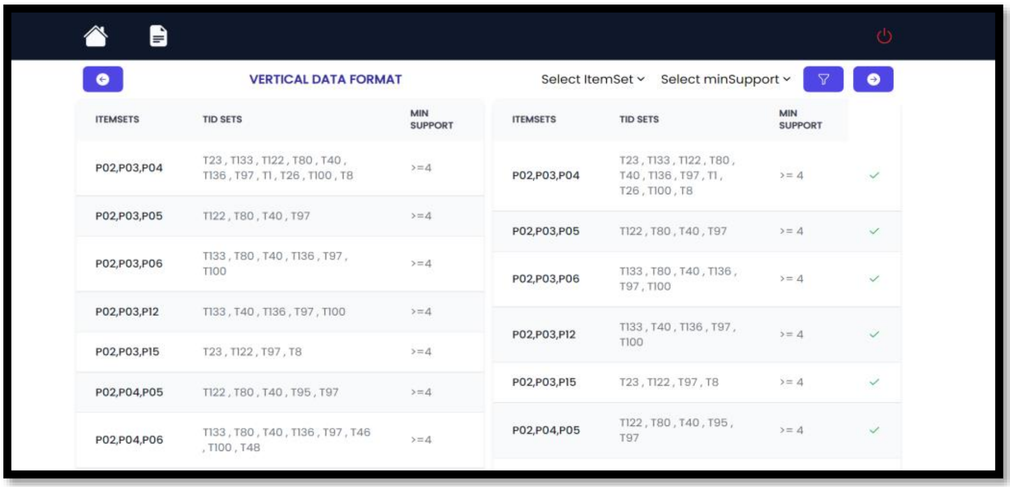
VERTICAL DATA FORMAT

ITEMSETS	TID SETS	MIN SUPPORT
P00	T107 , T20 , T133 , T11 , T51 , T98 , T19	>=2
P01	T97 , T9 , T5 , T34 , T134 , T10 , T23	>=2
P02	T36 , T9 , T23 , T133 , T43 , T122 , T57 , T50 , T86 , T63 , T59 , T68 , T124 , T80 , T40 , T107 , T67 , T95 , T136 , T111 , T97 , T60 , T117 , T1 , T121 , T125 , T85 , T46 , T115 , T18 , T112 , T74 , T82 , T89 , T113 , T62 , T19 , T69 , T26 , T49 , T100 , T27 , T128 , T58 , T65 , T72 , T108 , T35 , T92 , T8 , T88 , T38 , T87 , T94 , T41 , T48 , T52	>=2
P03	T11 , T10 , T23 , T26 , T122 , T28 , T59 , T100 , T21 , T27 , T63 , T30 , T97 , T1 , T131 , T96 , T40 , T8 , T80 , T4 , T136 , T29 , T133 , T7 , T39 , T44	>=2

ITEMSETS	TID SETS	MIN SUPPORT	
P00	T107 , T20 , T133 , T11 , T51 , T98 , T19	>= 2	✓
P01	T97 , T9 , T5 , T34 , T134 , T10 , T23	>= 2	✓
P02	T36 , T9 , T23 , T133 , T43 , T122 , T57 , T50 , T86 , T63 , T59 , T68 , T124 , T80 , T40 , T107 , T67 , T95 , T136 , T111 , T97 , T60 , T117 , T1 , T121 , T125 , T85 , T46 , T115 , T18 , T112 , T74 , T82 , T89 , T113 , T62 , T19 , T69 , T26 , T49 , T100 , T27 , T128 , T58 , T65 , T72 , T108 , T35 , T92 , T8 , T88 , T38 , T87 , T94 , T41 , T48 , T52	>= 2	✓

Figure 4.14 Display Vertically Format with Page List

In Figure 4.14, the system converts page list transactions to vertical format.



VERTICAL DATA FORMAT

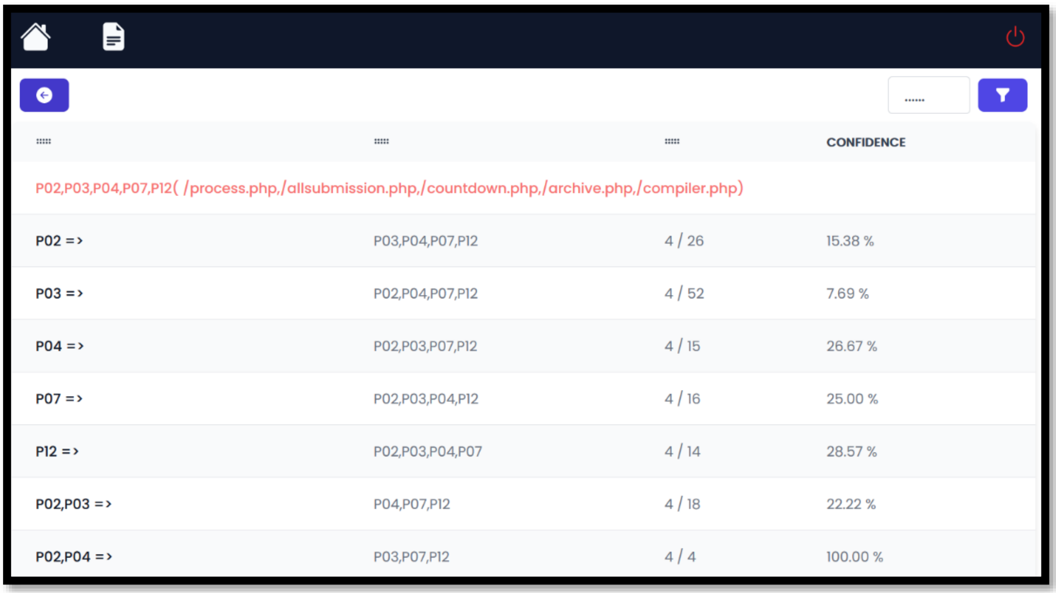
Select ItemSet ⌄ Select minSupport ⌄

ITEMSETS	TID SETS	MIN SUPPORT
P02,P03,P04	T23 , T133 , T122 , T80 , T40 , T136 , T97 , T1 , T26 , T100 , T8	>=4
P02,P03,P05	T122 , T80 , T40 , T97	>=4
P02,P03,P06	T133 , T80 , T40 , T136 , T97 , T100	>=4
P02,P03,P12	T133 , T40 , T136 , T97 , T100	>=4
P02,P03,P15	T23 , T122 , T97 , T8	>=4
P02,P04,P05	T122 , T80 , T40 , T95 , T97	>=4
P02,P04,P06	T133 , T80 , T40 , T136 , T97 , T46 , T100 , T48	>=4

ITEMSETS	TID SETS	MIN SUPPORT	
P02,P03,P04	T23 , T133 , T122 , T80 , T40 , T136 , T97 , T1 , T26 , T100 , T8	>= 4	✓
P02,P03,P05	T122 , T80 , T40 , T97	>= 4	✓
P02,P03,P06	T133 , T80 , T40 , T136 , T97 , T100	>= 4	✓
P02,P03,P12	T133 , T40 , T136 , T97 , T100	>= 4	✓
P02,P03,P15	T23 , T122 , T97 , T8	>= 4	✓
P02,P04,P05	T122 , T80 , T40 , T95 , T97	>= 4	✓

Figure 4.15 Displaying Paired Frequent Itemset

The end user can see page lists within the transactions set. And the end user can calculate frequent itemset, 2 itemset, 3 itemset, 4 itemset and 5 itemset by choosing mining support. The user can also filter frequent itemset by selecting itemset 1 to itemset 5. The system will display the final frequent itemset, 5 itemset.



......

::::	::::	::::	CONFIDENCE
P02,P03,P04,P07,P12(/process.php,/allsubmission.php,/countdown.php,/archive.php,/compiler.php)			
P02 =>	P03,P04,P07,P12	4 / 26	15.38 %
P03 =>	P02,P04,P07,P12	4 / 52	7.69 %
P04 =>	P02,P03,P07,P12	4 / 15	26.67 %
P07 =>	P02,P03,P04,P12	4 / 16	25.00 %
P12 =>	P02,P03,P04,P07	4 / 14	28.57 %
P02,P03 =>	P04,P07,P12	4 / 18	22.22 %
P02,P04 =>	P03,P07,P12	4 / 4	100.00 %

Figure 4.16 Calculate Confidence for Each Association rule

In this page, the system will represent the final result of frequent itemset or frequent webpages. The user can also see frequent pages together with page name. In this page, the user can view strong association rules. The system will calculate confidence for each frequent itemset because it wants to show more details of each of the page association. The system will display page association with percentage. The user can also filter percentage of the strong association rules.

CHAPTER 5

CONCLUSION AND FURTHER EXTENSION

The aim of the system is to generate interest frequent pattern on Online Judge web log record. Frequent pattern mining is also used to find information like set of pages repeatedly accessed together by web users. The administrator can modify the website according to the mining result as to restructure of webpages. In this paper, Online Judge of the University Server Log Dataset web user pattern discovery is discussed about. This work helps for the website user and administrator to improve the quality and performance. The website user easy to find the correct information to avoid irrelevant information and time wastage.

In this proposed system, to gain user interested pages or user behavior is the main target. In data pre-processing phase, data cleaning and defining the user pattern is most important fact. This system intends to help data cleaning and categorizing the web site for understanding the user interest and for improving the satisfactory of users' requirements. The contribution of the paper is to introduce the process of web log mining, and to show how frequent pattern discovery tasks can be applied on the web log data in order to obtain useful information about the user's navigation behaviors.

The proposed log record files which are very useful for each organization network. The data contains useful and meaningful information. This information is very difficult to understand. Frequent itemset calculation plays crucial role in most of the organization. Frequent itemset are those items which are frequently occurring in the database. A lot of algorithms have already been designed for generating frequent itemset. One of the algorithms that we have used is Eclat algorithm. In this system ECLAT algorithm is used to find frequent itemset and these itemset are calculated to be strong association rules by using confidence equation.

5.1 Advantages of the System

This system demonstrates effective use of association rule method on web log data. In this system describes the nature of step by step of preprocessing. It also used each algorithm for data cleaning, user identification, page identification and session identification. When finding the frequent itemset, the user can also use appropriate min_support. If the user does want to test other server log file, the system will accept the same format of web log file and generate frequent itemset for it. The system also calculate confidence for each frequent itemset to show strong association rules. The results obtained should absolutely help the website Analysts, Website Maintainers, Website Designers and Developers. It manages their system by analyzing occurred errors, corrupted and broken links.

5.2 Limitations of the System

Frequent itemset plays a very important role in our day to day life. Eclat algorithm is used to find frequent itemset. Eclat algorithm uses vertical dataset and bottom up approach for searching items in database. But in this proposed system has some limitations while finding frequent itemset. For example, large number of iterations is required for processing the items and more escape time is required for finding frequent itemset. So, the candidate five itemset are found and calculated confidence only the result of five frequent itemset. In the part of session identification, if the session time is longer, this system will delay when changing vertically data format. This system uses log record data from traditional log data one of the University’s website and not stream data flow.

5.3 Further Extensions

The implemented system can find frequent patterns by choosing support 2 or 3 or 4. To improve the association accuracy of the model, further studies should be conducted using different advanced association rule algorithms. And this system can be extended to test any web log file as in that program. And then this system can be used in other organization area such as e-commerce web log, online shopping web log. In the future, this proposed system is used to plan on different web sites with and to test the various web logs.

REFERENCES

1. Asfiya Khatoon1, Kuldeep Jaiswal, "*Web Page Ranking using Web Usage Mining*", International Journal of Advanced Research in Computer and Communication Engineering ISO 3297:2007 Certified Vol. 6, Issue 4, April 2017
2. Bina kotiyal,ankit kumar. Bhaskar pant, R.H. goudar, shivaji chauhan and sonam junee, "*User behavior analysis in web log through comparative study of Eclat and Apriori*" IEEE 2012
3. Dr. S. Vijayarani, Ms. P. Sathya "*An Efficient Algorithm for Mining Frequent Items in Data Streams*" in International Journal of Innovative Research in Computer and Communication Engineering Vol. 1, Issue 3, May 2013.
4. K.sudheer reddy, m. kantha reddy,v,sitaramula,"*An effective data preprocessing method for web usage mining*" IEEE conference 2013.
5. Kobra Etminani, Amin Rezaeian Delui,Noorali Raeeji Yanehsart, Modjtaba Rouhani. "*Web Usage Mining: Discovery Of The User's Navigational Patterns Using Som''*, IEEE 2009.
6. Mahendra Pratap Yadav, Mhd Feeroz, Vinod Kumar Yadav,"Mining The Customer Behavior Using Web Usage Mining In E-Commerce" IEEE 2012.
7. R. Krishnamoorthi, K.R. Suneetha, "*Extracting users pattern from web log data using decision tree and association rule*", Int. J. Business Performance and Supply Chain Modelling, Vol. 2, No. 2, 2010

8. S.VijayaKumar, 2A.S.Kumaresan, 3U.Jayalakshmi , "Frequent Pattern Mining in Web Log Data using Apriori Algorithm ." International Journal of Emerging Engineering Research and Technology Volume 3, Issue 10, October 2015, PP 50-55 .

9. Shamila Nasreen , "Frequent Pattern Mining Algorithms for Finding Associated Frequent Patterns for Data Streams, The 5th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN-2014).

10. Shaobo Shi; Yue Qi; Qin Wang, "FPGA Acceleration for Intersection Computation in Frequent Itemset Mining," Cyber-Enabled Distributed

Computing and Knowledge Discovery (CyberC), 2013 International Conference on , vol., no., pp.514,519, 10-12 Oct. 2013

11. Thanakorn Pamutha, Siriporn Chimphlee, Chom Kimpan, and Parinya Sanguansat, “Data Preprocessing on Web Server Log Files for Mining Users Access Patterns”, International Journal of Research and Reviews in Wireless Communications (IJRRWC), Vol. 2, No. 2, June 2012, ISSN: 2046-6447.

12. Yo unghee Kim, Won Young Kim and Ungmo Kim “*Mining frequent item sets with normalized weight in continuous data streams*”. Journal of information processing systems. 2010.

AUTHOR’S PUBLICATION

[1] Poe Myat Zin, Daw Aye Aye Maw, “*Frequent Patterrn Mining on Online Judge Education Web Log Data Using ECLAT Algorithm*”, in the Proceedings of the (Paper ID - 03019, Accepted Date:24 th June 2022) Conference on Parallel and Soft Computing (PSC 2022), Yangon, Myanmar, 2022.